Accelerated Systems for Precision Health

EECS 570
Satish Narayansamy

Ack: Reetu Das
What is Precision Health?

“an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person”

“Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?”

- President Obama, January 30, 2015
What is Precision Health?

One Size Fits All  X

Precision Medicine  ✓
Precision Health Platform

Data Sources
- Electronic Health Records
- Genome
- Epigenome
- Microbiome
- Metabolome

Decision Support Tools
- Big Data Analytics
- Sequencing
- Analysis
- Big data analytics + machine learning models + db

Clinical Decisions
- Diagnosis and Prevention
- Treatment and Management

Credits: Created from BioRender.com
System Design Considerations

- Efficiency
- Performance
- Latency-critical
- Throughput-oriented
- Energy
- Cost
- Security
- Form factor
- Cloud
- Edge

Efficiency

Security and Privacy

Form Factor

Encrypt(VCF_1, key_1)
Encrypt(VCF_2, key_2)
Encrypt(VCF_3, key_3)

Homomorphic encryption, Intel SGX

Credits: Created from BioRender.com
Sequencing is Key Ingredient of Precision Health
Exponential Growth in Genome Sequencing

Credits: [Stephens et al. PLOS Bio, 2015] [Illumina] [Oxford Nanopore] [10x Chromium][Biorender.com]
Sequencing Costs have Plummeted
Exploding Sequencing Applications

- Cancer treatment
- Microbiome
- In operating room sequencing
- Liquid Biopsy
- Consumer Genotyping
- Agricultural sequencing
- Portable Pathogen detector
- Food Safety
**Sequencing Technologies: Evolution**

**Illumina Sequencing by Synthesis**
- **Illumina Genome Analyzer, 2005**
- **Illumina NovaSeq 6000, 2021**
- **1 Gbases/per day**
- **3 Tbases/per day**
- **Read length:** 100-350bp
- **Per base inaccuracy:** 0.1%
- **1000x increase in sequencing machine throughput**

**Nanopore Sequencing**
- **Read length:** 1kb-1Mbp
- **Per base inaccuracy:** 1-15%
- **1000x increase in sequencing fragment length**
- **10 - 100x increase in sequencing error rate**

Credits: Illumina, DataBase Center for Life Science (DBCLS), [https://doi.org/10.7875/togopic.2020.01](https://doi.org/10.7875/togopic.2020.01), Wikipedia DMLapato
[https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina](https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina)
Nanopore Sequencing is poised to revolutionize molecular diagnostics

- Nanopore sequencing feeds DNA strands through a biological pore in a membrane
- Current disruptions across the membrane are recorded
- Current disruptions correspond to individual DNA base-pairs (A, T, G, C)

- Thousands of parallel pores are embedded into a “flowcell”
- Flowcells are run via a hand-held, USB-powered device called a MinION
Nanopore Sequencing is poised to revolutionize molecular diagnostics
Key advantages versus traditional Next Generation Sequencing technologies:

- Long reads (>4Mb record)
- Simple library preparation
- Streaming output
- Small and portable
- RNA + DNA sequencing
- Low cost ($1k)
Nanopore Sequencing Lab at UM EECS

- Biosafety Level -2 Certification for tissue and RNA work
- Standard molecular biology equipment
- Small -20C freezer
- Enables tight coupling of informatics with nanopore sequencer
End-to-End Application Case Studies
Acceleration Study – Whole Genome Sequencing
Acceleration Study: Whole Genome Sequencing

Human Genome
6 G bases

Sequenced reads (~billions)

Reference genome

Read

Aligned reads

Read Alignment

Reference genome

Duplicate Marking

Sorting +

Variant Calling

Diagnosis

BWA-MEM

Picard SortSam,
MarkDuplicates

GATK Haplotype Caller

Baseline
m5.8xlarge
32 vCPUs

13.9 hr (i.e. 445 CPU hrs)
SillaX ASIC fabricated (55nm) 63x faster than 56-thread CPU SeqAn for 100bp reads
Read Alignment: SeedEx

Full-band implementation

- Low utilization
- Area inefficient

Banded implementation

- Miss Optimal path

Speculatively compute with a narrow band PE array

SeedEx check algorithm
Uses admissive heuristics. If optimality cannot be guaranteed, fall back to CPU/full-band machine

Accuracy

100% equivalent results on AWS cloud FPGA when integrated with BWA-MEM software

2.3x smaller than banded Smith Waterman core (w = 41 + edit machine)

6x higher throughput over banded Smith-Waterman FPGA (w = 101) for same area
Read Alignment: ERT

Problem

- Seeding
- Reference
- Seeds

Our Solution

- Bandwidth-efficient data structure: Enumerated Radix Tree (ERT)
- Bandwidth-efficient search algorithm

Results

- 2.3x over BWA-MEM2 with SeedEx
- Open-source: https://github.com/bwa-mem2/bwa-mem2/tree/ert

- FM-index → widely used seeding data structure
- Memory-bandwidth bottleneck

- FM-index 4.2 GB human

- ERT ~60 GB human

- Trades-off memory capacity for memory bandwidth to improve seeding performance
- Supports multi-character lookup with index table and customized radix tree. Used to implement optimized longest match search algorithm in BWA-MEM
ERT software integration with Broad Institute / Intel’s BWA-MEM 2

BWA-MEM is the de-facto genomics read alignment tool used by researchers and practitioners worldwide
### Sorting/Duplicate Marking Optimizations

- I/O bandwidth bound. Optimized counting sort based multi-thread CPU implementation

- Same results as Picard SortSam and Picard MarkDuplicates

- Runtime: +3 min for 50x coverage WGS alignments (56 thread CPU)
  Memory: ~75 GB memory
Variant Calling: pairHMM Acceleration

Pruning Algorithm

Haplotype

A

T

G

C

A

Read

Result = sum of last row

Squares need to be processed by floating point machine

Squares processed by pruning machine

Dominated by this alignment

Acceleraotr Architecture

On-demand Job Scheduler

Reduce idle machines due to work load imbalance between pruning machine & precise machine

Read Regfile

Haplotype Regfile

Intra Job Parallelism

Precise Machine (Floating Point)

Input Regfile

Haplotype Regfile

Unpruned Cells

Floating Point PE

Output lower bound

Output upperbound

Unpruned cells

Pruning pairHMM ASIC (40nm)

Memorv

“Refinement”

“Scan” machines

Arbiter

“Scan” machines

Bit equivalent output

43x fewer cells computed in precise floating point

8.3x higher throughput (GCUPS) than floating-point ASIC of the same area
Why Accuracy Matters?

- Human ~ Chimpanzee 96%
- Human ~ Cat 90%
- Human ~ Cow 80%
- Human ~ Banana 50-60%

Slide credit: Onur Mutlu, "Accelerating Genome Analysis: A Primer on an Ongoing Journey"
Demo: Aligning reads from CYP2C19 gene
Effect of G->A variant in the CYP2C19 gene

CYP2C19 involved in metabolism of > 10% commonly prescribed drugs

Normal

DNA: GT AG CCGGG

Transcription

GT-AG Splicing

Normal RNA

Translation

Protein: ☑

Non-coding (intron)

Aberrant

DNA: GT AG CCCAGG

Transcription

GT-AG Splicing

Aberrant RNA

Translation

Aberrant Protein: ☒

Coding (exon)

40-bp deletion
**Acceleration Study: Whole Genome Sequencing**

- **Baseline**
  - m5.8xlarge
  - 32 vCPUs
  - Time: 13.9 hr

- **FPGA system**
  - f1.4xlarge
  - + 16 vCPUs
  - Time: 2.5 hr
  - **~5.5x**

- **ASIC system**
  - 1.9 hr
  - **~7.3x**

---

**Other Acceleration Candidates**

- **Chaining**
  - FPGA system
  - f1.4xlarge
  - + 16 vCPUs
  - Time: 13.9 hr
  - **~5.5x**

- **De-Bruijn graph assembly**
  - ASIC system
  - 1.9 hr
  - **~7.3x**

---

**Active Region of Reference**

- **Hash Table**
  - CACT
  - TGAA
  - ACTC
  - CCAC
  - CTGC
  - TGA
  - ACTG

- **De-Bruijn graph**
  - CACT
  - ACTC
  - CCAC
  - CTGC
  - TGA
  - TCAA
  - GAAC

---

**Seeding**

- **Sorting / Mark Duplicates**
- **pairHMM**
- **Other**

---

**Time (hr)**

- 0
- 5
- 10
- 15
Acceleration Study – Ultra Rapid Cancer Diagnosis
Intra-operative sequencing for accurate cancer diagnostics

- Intra-operative histology can help guide surgical decision making and combine surgeries.
- Histology is subjective, and does not contain molecular information.
- Genetic information is becoming increasingly important for diagnosis and targeted, personalized treatment!

Frozen Section Histology can return a diagnosis in ~20-40 min.

“For the first time, the WHO classification of CNS tumors uses molecular parameters in addition to histology to define many tumor entities, thus formulating a concept for how CNS tumor diagnoses should be structured in the molecular era.”

Can we sequence a tumor’s DNA within the intra-operative time frame? (i.e. <1hr)
How does a sequencing-based molecular diagnostic work?

- Target amplification uses the Polymerase Chain Reaction (PCR) to exponentially amplify a region of the genome.
- PCR exponentially amplifies a small cancer-relevant gene target that might contain a mutation.
- Amplified targets can then be sequenced to determine if a mutation is present.

Target amplification is the obvious bottleneck. How can we attack this?
Threshold Sequencing

Co-optimize amplification time and sequencing time to minimize time-to-result

1) Build a model to estimate total diagnostic time

\[ T_{total} = T_{amp} + T_{seq} \]

\[ T_{amp} = T_{init} + T_{cycle} \times N_{cycle} + T_{final} \]

\[ F_{target} = \frac{2^{N_{cycle}}}{2^{N_{cycle}} + N_{background}} \]

\[ T_{seq} = N_{depth} \times \frac{1}{N_{pores} \times R_{sample} \times F_{target}} \]

2) Augment model with experimentally derived parameters

3) Run diagnostic with final optimal parameters

Co-optimization allowed for a world-first demonstration of a sub-1 hour sequencing-based diagnostic

DNA Extraction 10 min
Target Amplification 26 min
Library Preparation 11 min
Sequencing 5 min
Informatics Real-Time

52 min Diagnostic

but target amplification is still a large bottleneck...
Loop-Mediated Isothermal Amplification (LAMP) Technology

**Benefits**
- LAMP amplifies targets much more rapidly than PCR (14min vs 26min)
- LAMP generates concatemeric reads that contain redundant, and complementary information

** Downsides**
- Difficult to analyze and reason about complex product
- No LAMP specific bioinformatics tools

We leverage LAMP’s rapid amplification and redundant information to further reduce diagnostic time

https://doi.org/10.1016/j.trac.2019.01.015

N=6 concatemer

N=1 target
LAMPrey: a new bioinformatics tool to analyze and “polish” LAMP concatemer product

LAMPrey identifies concatemer “sub-reads” in noisy amplicons

LAMPrey is able to recover about 50% more information than traditional informatics tools

Information from each sub-read can be combined to form a more confident base call (polishing) resulting in a more rapid and accurate diagnostic
LAMPrey + Threshold Sequencing = <30min Sequencing-based Diagnostic

Experimentally informed LAMP diagnostic model

Final LAMP diagnostic result

LAMPrey benefit

LAMPrey and other optimizations allowed for a world-first demonstration of a sub-30 minute sequencing-based diagnostic

DNA Extraction 5.5 min
Target Amplification 15 min
Library Preparation 5 min
Sequencing 3.5 min
Informatics Real-Time

<30 min Diagnostic

Open source: https://www.github.com/jackwadden/lamprey
LAMPprey + Threshold Sequencing = <30min Sequencing-based Diagnostic

Sets world record for fastest time-to-result

DNA Extraction: 5.5 min
Target Amplification: 15 min
Library Preparation: 5 min
Sequencing: 3.5 min
Informatics: Real-Time

LAMPprey and other optimizations allowed for a world-first demonstration of a sub-30 minute sequencing-based diagnostic
Acceleration Study – Real-Time Pathogen Detection
Viral Pandemics & Rise of Superbugs

Coronavirus Cases: 30,862,212
Deaths: 561,225

Coronaviruses
SARS, MERS and 2019-nCoV

Superbugs will kill more than cancer by 2050

2019 UN report: “No Time To Wait”
It Took Months For Mass COVID Testing Capabilities

PCR Primer Design Complexity
How Can We Be Ready For The Next Pandemic?

Universal Pathogen Detector

Digitally programmable using target pathogen’s genome

distribute detectors

when novel pathogen is identified

digitally distribute pathogen’s genome
Advantages

► Real-time, portable, low-latency

► Digital primers that can dispersed worldwide instantaneously
  ➤ No need for complex PCR primer development & deployment delaying diagnosis
  ➤ Rapid early stage diagnosis to avoid pandemic

► Possible to detect “unknown” pathogens
  ➤ Primer based molecular tests can detect only known pathogens
  ➤ Genetic “sniffers” to act before start of pandemic

► Strain typing for tracking disease progression
  ➤ Complete molecular sequence of pathogen available
  ➤ Phylogenic evolutionary relationships can be analyzed
Pathogen detection pipeline: Our Approach

Specimen acquisition

DNA extraction

Basecaller

Accelerated Filter

Sequence analysis

Interpretation

Strengths

Access to 1000s of samples to conduct clinical trials

Translational Medicine:
Expertise in infectious diseases and sequencing

Computer Hardware and Systems:
Expertise to develop hardware accelerated bioinformatics pipelines

Machine Learning:
Expertise to develop machine learning methods
Pathogen detection pipeline: Our Approach

Specimen acquisition

DNA extraction

Skip useless reads (“read-until”)

Basecaller

Accelerated Filter

Sequence analysis

Stop: data sufficient (“read-until”)

Solutions developed

~20 min using off-the-shelf kits

Achieved world’s first real-time metagenomics demonstration for diagnosing pneumonia - Dickson Lab (collaborator)

Can filter 99% of human reads

Save time and cost by skipping human reads (read-until)

Raw signal to host classifier

Read alignment accelerator

Stop when data is sufficient (leverage ONT’s real-time feature)
From nasal swab to SARS-CoV-2 genome in under 2 hours

Read Until helps reduce sequencing time and cost by 38%

PathFinder compared to existing Read Until pipeline on high-end GPU (Titan XP):

456x lower latency  3,283x higher throughput/Watt

3,109x higher throughput/area
Demonstration of Pathogen Abundance Estimation (unbiased)

Time: ~1 hr
10 min. DNA extraction time
10 min. library prep time
40 min. sequencing + compute analysis

Cost: ~$90

Primary analysis sequencing using PromethION

Input Sample
Human:Bacteria read ratio: 99:1

Correctly filtered all human reads

Real AE output
Our AE output
How Can You Kick-Start Precision Health Research?
GenomicsBench

Open-source:

https://github.com/arun-sub/genomicsbench

12 computationally intensive kernels drawn from well maintained software tools

Covers the major steps of modern sequence analysis pipelines

Includes both short and long read analysis algorithms

Small/large input datasets
“Discover the genetic, lifestyle and environmental factors that influence a population’s health and provides personalized solutions that allow individuals to improve their health and wellness.”
Work from Awesome Group of Fantastic Students!!

“Discover the genetic, lifestyle and environmental factors that influence a population’s health and provides personalized solutions that allow individuals to improve their health and wellness.”